

# Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees<sup>\*†</sup>

Matteo Riondato<sup>‡</sup> and Eli Upfal

Department of Computer Science, Brown University, Providence, RI, USA  
{matteo, eli}@cs.brown.edu

June 22, 2012

## Abstract

The tasks of extracting (top- $K$ ) Frequent Itemsets (FI's) and Association Rules (AR's) are fundamental primitives in data mining and database applications. Exact algorithms for these problems exist and are widely used, but their running time is hindered by the need of scanning the entire dataset, possibly multiple times. High quality approximations of FI's and AR's are sufficient for most practical uses, and a number of recent works explored the application of sampling for fast discovery of approximate solutions to the problems. However, these works do not provide satisfactory performance guarantees on the quality of the approximation, due to the difficulty of bounding the probability of under- or over-sampling any one of an unknown number of frequent itemsets. In this work we circumvent this issue by applying the statistical concept of *Vapnik-Chervonenkis (VC) dimension* to develop a novel technique for providing tight bounds on the sample size that guarantees approximation within user-specified parameters. Our technique applies both to absolute and to relative approximations of (top- $K$ ) FI's and AR's. The resulting sample size is linearly dependent on the VC-dimension of a range space associated with the dataset to be mined. The main theoretical contribution of this work is a characterization of the VC-dimension of this range space and a proof that it is upper bounded by an easy-to-compute characteristic quantity of the dataset which we call *d-index*, namely the maximum integer  $d$  such that the dataset contains at least  $d$  transactions of length at least  $d$ . We show that this bound is strict for a large class of datasets. The resulting sample size for an absolute (resp. relative)  $(\epsilon, \delta)$ -approximation of the collection of FI's is  $O(\frac{1}{\epsilon^2}(d + \log \frac{1}{\delta}))$  (resp.  $O(\frac{2+\epsilon}{\epsilon^2(2-\epsilon)\theta}(d \log \frac{2+\epsilon}{(2-\epsilon)\theta} + \log \frac{1}{\delta}))$ ) transactions, which is a significant improvement over previous known results. We present an extensive experimental evaluation of our technique on real and artificial datasets, demonstrating the practicality of our methods, and showing that they achieve even higher quality approximations than what is guaranteed by the analysis.

## 1 Introduction

Discovery of frequent itemsets and association rules is a fundamental computational primitive with application in data mining (market basket analysis), databases (histogram construction), networking (heavy hitters) and more [15, Sect. 5]. Depending on the particular application, one is interested in finding all itemsets with frequency greater or equal to a user defined threshold (FIs), identifying the  $K$  most frequent itemsets (top- $K$ ), or computing all association rules (ARs) with user defined minimum support and confidence level. Exact solutions to these problems require scanning the entire dataset, possibly multiple times. For large datasets that do not fit in main memory, this can be prohibitively expensive. Furthermore, such extensive computation is often unnecessary, since high quality approximation are sufficient for most practical applications. Indeed, a number of recent papers [4, 6, 7, 9, 10, 12, 13, 17–22, 26, 28, 30–34, 37–42] explored the application of sampling for approximate solutions to these problems. However, the efficiency and practicality of the sampling approach depends on a tight relation between the size of the sample and the quality of the resulting approximation. Previous works do not provide satisfactory solutions to this problem.

The technical difficulty in analyzing any sampling technique for frequent itemsets discovery problems is that a-priori any subset of items can be among the most frequent ones, and the number of subsets is exponential in the number of distinct items appearing in the dataset. A standard analysis begins with a bound on the probability that a given itemset is either over or under represented in the sample. Such bound is easy to obtain using a Chernoff-like bound or the Central Limit theorem. The difficulty is in combining the bounds for individual itemsets into a global bound that holds simultaneously for all the itemsets. A simple application of the union bound vastly overestimates the error probability because of the large number of possible itemsets, a large fraction of which may not be present in the dataset and therefore should not be considered. More sophisticated techniques, developed in recent works [6, 12, 31],

<sup>\*</sup>Work was supported in part by NSF award IIS-0905553.

<sup>†</sup>A shorter version of this paper is scheduled to appear in the proceedings of ECML PKDD 2012.

<sup>‡</sup>Contact author.

give better bounds only in limited cases. A loose bound on the required sample size for achieving the user defined performance guarantees, decreases the gain obtained from the use of sampling.

In this work we circumvent this problem through a novel application of the *Vapnik-Chervonenkis* (VC) dimension concept, a fundamental tool in statistical learning theory. Roughly speaking, the VC-dimension of a collection of indicator functions (a range space) is a measure of its complexity or expressiveness (see Sect. 2.2 for formal definitions). A major result [36] relates the VC-dimension of a range space to the sufficient size for a random sample to simultaneously approximate all the indicator functions within predefined parameters. The main obstacle in applying the VC-dimension theory to particular computation problems is computing the VC-dimension of the range spaces associated with these problems.

We apply the VC-dimension theory to frequent itemsets problems by viewing the presence of an itemset in a transaction as the outcome of an indicator function associated with the itemset. The major theoretical contributions of our work are a complete characterization of the VC-dimension of the range space associated with a dataset, and a tight bound to this quantity. We prove that the VC-dimension is upper bounded by an easy-to-compute characteristic quantity of the dataset which we call *d-index*, namely, the maximum integer  $d$  such that the dataset contains at least  $d$  transactions of length at least  $d$ . We show that this bound is tight by demonstrating a large class of datasets with a VC-dimension that matches the bound.

The VC-dimension approach provides a unified tool for analyzing the various frequent itemsets and association rules problems (i.e., the market basket analysis tasks). We use it to prove tight bounds on the required sample size for extracting FI's with a minimum frequency threshold, for mining the top- $K$  FI's, and for computing the collection of AR's with minimum frequency and confidence thresholds. Furthermore, we compute bounds for both absolute and relative approximations (see Sec 2.1 for definitions). We show that high quality approximations can be obtained by mining a very small random sample of the dataset. For example, the required sample size for an absolute  $(\varepsilon, \delta)$ -approximation of the collection of FI's is  $O(\frac{1}{\varepsilon^2}(d + \log \frac{1}{\delta}))$  transactions, which is a significant improvement over previous known results, as it is smaller and, more importantly, less dependent on parameters such as the minimum frequency threshold and the dataset size. Similar results are proven for the top- $K$  FI's and AR's tasks.

We present an extensive experimental evaluation of our method using real and artificial datasets, to assess the practicality of our approach. The experimental results show that indeed our method achieves, and even exceeds, the analytically proven guarantees for the quality of the approximations.

## 1.1 Previous Work

Agrawal et al. [1] introduced the problem of mining association rules in the basket data model, formalizing a fundamental task of information extraction in large datasets. Almost any known algorithm for the problem starts by solving a FI's problem and then generate the association rules implied by these frequent itemsets. Agrawal and Srikant [2] presented *Apriori*, the most well-known algorithm for mining FI's, and *FastGenRules* for computing association rules from a set of itemsets. Various ideas for improving the efficiency of FI's and AR's algorithms have been studied, and we refer the reader to the survey by Ceglar and Roddick [5] for a good presentation of recent contributions. However, the running times of all known algorithms heavily depend on the size of the dataset.

Mannila et al. [28] first suggested the idea that sampling can be used to efficiently obtain the collection of FI's, presenting some empirical results to validate the intuition. Toivonen [34] presents an algorithm that, by mining a random sample of the dataset, builds a candidate set of frequent itemsets which contains all the frequent itemsets with a probability that depends on the sample size. There are no guarantees that all itemsets in the candidate set are frequent, but the set of candidates can be used to efficiently identify the set of frequent itemsets with at most two passes over the entire dataset. The work also suggests a bound on the sample size sufficient to ensure that the frequencies of itemsets in the sample are close to their real one. The analysis uses Chernoff bounds and the union bound. The major drawback of this sample size is that it depends linearly on the number of individual items appearing in the dataset.

Zaki et al. [40] show that static sampling is an efficient way to mine a dataset, but choosing the sample size using Chernoff bounds is too conservative, in the sense that it is possible to obtain the same accuracy and confidence in the approximate results at smaller sizes than what the theoretical analysis suggested.

Other works tried to improve the bound to the sample size by using different techniques from statistic and probability theory like the central limit theorem [19, 22, 41] or hybrid Chernoff bounds [42].

Since theoretically-derived bounds to the sample size were too loose to be useful, a corpus of works applied progressive sampling to extract FI's [4, 7, 9, 10, 12, 17, 18, 20, 21, 26, 30, 39]. Progressive sampling algorithms work by selecting a random sample and then trimming or enriching it by removing or adding new sampled transactions according to a heuristic or a self-similarity measure that is fast to evaluate, until a suitable stopping condition is satisfied. The major downside of this approach is that it offers no guarantees on the quality of the obtained results.

Another approach to estimating the required sample size is presented in [13]. The authors give an algorithm that studies the distribution of frequencies of the itemsets and uses this information to fix a sample size for mining frequent itemsets, but without offering any theoretical guarantee.

A recent work by Chakaravarthy et al. [6] gives the first analytical bound on a sample size that is linear in the length of the longest transaction, rather than in the number of items in the dataset. This work is also the first to present an algorithm that uses a random

sample of the dataset to mine approximated solutions to the AR's problem with quality guarantees. No experimental evaluation of their methods is presented, and they do not address the top- $K$  FI's problem. Our approach gives better bounds for the problems studied in [6] and applies to related problems such as the discovery of top- $K$  FI's and absolute approximations.

Extracting the collection of top- $K$  frequent itemsets is a more difficult task since the corresponding minimum frequency threshold is not known in advance [11, 14]. Some works solved the problem by looking at *closed* top- $K$  frequent itemsets, a concise representation of the collection [32, 38], but they suffers from the same scalability problems as the algorithms for exactly mining FI's with a fixed minimum frequency threshold.

Previous works that used sampling to approximation the collection of top- $K$  FI's [31, 33] used progressive sampling. Both works provide (similar) theoretical guarantees on the quality of the approximation. What is more interesting to us, both works present a theoretical upper bound to the sample size needed to compute such an approximation. The size depended linearly on the number of items. In contrast, our results give a sample size that only in the worst case is linear in the number of items but can be (and is, in practical cases) much less than that, depending on the dataset, a flexibility not provided by previous contributions. Sampling is used by Vasudevan and Vojonović [37] to extract an approximation of the top- $K$  frequent individual *items* from a sequence of items, which contains no item whose actual frequency is less than  $f_K - \varepsilon$  for a fixed  $0 < \varepsilon < 1$ , where  $f_K$  is the *actual* frequency of the  $K$ -th most frequent item. They derive a sample size sufficient to achieve this result, but they assume the knowledge of  $f_K$ , which is rarely the case. An empirical sequential method can be used to estimate the right sample size. Moreover, the results cannot be directly extended to the mining of top- $K$  frequent item(set)s from datasets of transactions with length greater than one.

## 1.2 Our Contributions

By applying tools from statistical learning theory, we develop a general technique for bounding the sample size required for generating high quality approximations to frequent itemsets and association rules tasks. Table 1 compares our technique to the best previously known results for the various problems (see Sect. 2.1 for definitions). Our bounds, which are linear in the VC-dimension associated with the dataset, are consistently smaller and less dependent on other parameters of the problem than previous results. An extensive experimental evaluation demonstrates the advantage of our technique in practice.

Task	Approximation	This work	Best previous work
FI's	absolute	$\frac{4c}{\varepsilon^2} (d + \log \frac{1}{\delta})$	$O(\frac{1}{\varepsilon^2} ( \mathcal{I}  + \log \frac{1}{\delta}))$ [19, 22, 34, 41]
	relative	$\frac{4(2+\varepsilon)c}{\varepsilon^2(2-\varepsilon)\theta} (d \log \frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log \frac{1}{\delta})$	$\frac{24}{\varepsilon^2(1-\varepsilon)\theta} (\Delta + 5 + \log \frac{4}{(1-\varepsilon)\theta\delta})$ [6]
top- $K$ FI's	absolute	$\frac{16c}{\varepsilon^2} (d + \log \frac{1}{\delta})$	$O(\frac{1}{\varepsilon^2} ( \mathcal{I}  + \log \frac{1}{\delta}))$ [31, 33]
	relative	$\frac{4(2+\varepsilon)c}{\varepsilon^2(2-\varepsilon)\theta} (d \log \frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log \frac{1}{\delta})$	not available
AR's	absolute	$O(\frac{(1+\varepsilon)}{\varepsilon^2(1-\varepsilon)\theta} (d \log \frac{1+\varepsilon}{\theta(1-\varepsilon)} + \log \frac{1}{\delta}))$	not available
	relative	$\frac{16c(4+\varepsilon)}{\varepsilon^2(4-\varepsilon)\theta} (d \log \frac{4+\varepsilon}{\theta(4-\varepsilon)} + \log \frac{1}{\delta})$	$\frac{48}{\varepsilon^2(1-\varepsilon)\theta} (\Delta + 5 + \log \frac{4}{(1-\varepsilon)\theta\delta})$ [6]

Table 1: Required sample sizes (as number of transactions) as a function of the VC-dimension  $d$ , the maximum transaction size  $\Delta$ , the number of items  $|\mathcal{I}|$ , the accuracy  $\varepsilon$ , the failure probability  $\delta$ , the minimum frequency  $\theta$ , and the minimum confidence  $\gamma$ . Note that  $d \leq \Delta \leq |\mathcal{I}|$ .

To the best of our knowledge, our work is the first to provide a characterization and an explicit bound for the VC-dimension of the range space associated to a dataset and to apply the result to the extraction of FI's and AR's from random sample of the dataset. We believe that this connection with statistical learning theory can be furtherly exploited in other data mining problems.

We also believe that our approach can be applied not just to mining collections of frequent itemsets and association rules, which can be massive, but also to the mining of small collections of itemsets/association rules that describe the dataset with the minimal number of itemsets/association rules possible, as presented in [27].

**Outline.** In Sect. 2 we formally define the problem and our goals, and introduce definitions and lemmas used in the analysis. The main part of the analysis with derivation of a strict bound to the VC-dimension of association rules is presented in Sect. 3, while our algorithms and sample sizes for mining FI's, top- $K$  FI's, and association rules through sampling are in Sect. 4. Section 5 contains an extensive experimental evaluation of our techniques.

## 2 Preliminaries

We now introduce basic definitions and lemmas we will use in later sections.

## 2.1 Datasets, Itemsets, and Association Rules

A *dataset*  $\mathcal{D}$  is a collection of *transactions*, where each transaction  $\tau$  is a subset of a ground set  $\mathcal{I}$ . There can be multiple identical transactions in  $\mathcal{D}$ . Members of  $\mathcal{I}$  are called *items* and members of  $2^{\mathcal{I}}$  are called *itemsets*. Let  $|\tau|$  denote the number of items in transaction  $\tau$ . Given an itemset  $A \in 2^{\mathcal{I}}$ , let  $T_{\mathcal{D}}(A)$  denote the set of transactions in  $\mathcal{D}$  that contain  $A$ . The *support* of  $A$ ,  $\sigma_{\mathcal{D}}(A) = |T_{\mathcal{D}}(A)|$ , is the number of transaction in  $\mathcal{D}$  that contains  $A$ , and the *frequency* of  $A$ ,  $f_{\mathcal{D}}(A) = \frac{|T_{\mathcal{D}}(A)|}{|\mathcal{D}|}$ , is the fraction of transactions in  $\mathcal{D}$  that contain  $A$ .

**Definition 1.** Given a *minimum frequency threshold*  $\theta$ ,  $0 < \theta \leq 1$ , the *FI's mining task with respect to  $\theta$*  is finding all itemsets with frequency  $\geq \theta$ , i.e., the set

$$\text{FI}(\mathcal{D}, \mathcal{I}, \theta) = \{(A, f_{\mathcal{D}}(A)) : A \in 2^{\mathcal{I}} \text{ and } f_{\mathcal{D}}(A) \geq \theta\}.$$

To define the collection of top- $K$  FI's, we assume a fixed *canonical ordering* of the itemsets in  $2^{\mathcal{I}}$  by decreasing frequency in  $\mathcal{D}$ , with ties broken arbitrarily, and label the itemsets  $A_1, A_2, \dots, A_m$  according to this ordering. For a given  $K$ , with  $1 \leq K \leq m$ , we denote with  $f_{\mathcal{D}}^{(K)}$  the frequency  $f_{\mathcal{D}}(A_K)$  of the  $K$ -th most frequent itemset  $A_K$ , and define the set of top- $K$  FI's (with their respective frequencies) as

$$\text{TOPK}(\mathcal{D}, \mathcal{I}, K) = \text{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}).$$

One of the main uses of frequent itemsets is in the discovery of *association rules*.

**Definition 2.** An *association rule*  $W$  is an expression " $A \Rightarrow B$ " where  $A$  and  $B$  are itemsets such that  $A \cap B = \emptyset$ . The *support*  $\sigma_{\mathcal{D}}(W)$  (resp. frequency  $f_{\mathcal{D}}(W)$ ) of the association rule  $W$  is the support (resp. frequency) of the itemset  $A \cup B$ . The *confidence*  $c_{\mathcal{D}}(W)$  of  $W$  is the ratio  $\frac{f_{\mathcal{D}}(A \cup B)}{f_{\mathcal{D}}(A)}$  of the frequency of  $A \cup B$  to the frequency of  $A$ .

Intuitively, an association rule " $A \Rightarrow B$ " expresses, through its support and confidence, how likely it is for the itemset  $B$  to appear in the same transactions as itemset  $A$ , so that when  $A$  is found in a transaction it is then possible to infer that  $B$  will be present in the same transaction with a probability equal to the confidence of the association rule.

**Definition 3.** Given a dataset  $\mathcal{D}$  with transactions built on a ground set  $\mathcal{I}$ , and given a minimum frequency threshold  $\theta$  and a minimum confidence threshold  $\gamma$ , the *AR's task with respect to  $\theta$  and  $\gamma$*  consist in finding the set

$$\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma) = \{(W, f_{\mathcal{D}}(W), c_{\mathcal{D}}(W)) \mid \text{Association Rule } W, f_{\mathcal{D}}(W) \geq \theta, c_{\mathcal{D}}(W) \geq \gamma\}.$$

Often, with an abuse of the notation, we will say that an itemset  $A$  (resp. an association rule  $W$ ) is in  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  or in  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  (resp. in  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ ) and denote this fact with  $A \in \text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  or  $A \in \text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  (resp.  $W \in \text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ ), meaning that there is a pair  $(A, f) \in \text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  or  $(A, f) \in \text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  (resp. a triplet  $(W, f_w, c_w) \in \text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ ).

In this work we are interested in extracting absolute and relative approximations of the sets  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ ,  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  and  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ .

**Definition 4.** Given a parameter  $\varepsilon_{\text{abs}}$  (resp.  $\varepsilon_{\text{rel}}$ ), an *absolute  $\varepsilon_{\text{abs}}$ -close approximation* (resp. a *relative  $\varepsilon_{\text{rel}}$ -close approximation*) of  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  is a set  $\mathcal{C} = \{(A, f_A) : A \in 2^{\mathcal{I}}, f_A \in [0, 1]\}$  of pairs  $(A, f_A)$  where  $f_A$  approximates  $f_{\mathcal{D}}(A)$ .  $\mathcal{C}$  is such that:

1.  $\mathcal{C}$  contains all itemsets appearing in  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ ;
2.  $\mathcal{C}$  contains no itemset  $A$  with frequency  $f_{\mathcal{D}}(A) < \theta - \varepsilon_{\text{abs}}$  (resp.  $f_{\mathcal{D}}(A) < (1 - \varepsilon_{\text{rel}})\theta$ );
3. For every pair  $(A, f_A) \in \mathcal{C}$ , it holds  $|f_{\mathcal{D}}(A) - f_A| \leq \varepsilon_{\text{abs}}$  (resp.  $|f_{\mathcal{D}}(A) - f_A| \leq \varepsilon_{\text{rel}} f_{\mathcal{D}}(A)$ ).

This definition extends easily to the case of top- $K$  frequent itemsets mining using the equivalence

$$\text{TOPK}(\mathcal{D}, \mathcal{I}, K) = \text{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}) :$$

an absolute (resp. relative)  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)})$  is an absolute (resp. relative)  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ .

For the case of association rules, we have the following definition.

**Definition 5.** Given a parameter  $\varepsilon_{\text{abs}}$  (resp.  $\varepsilon_{\text{rel}}$ ), an *absolute  $\varepsilon_{\text{abs}}$ -close approximation* (resp. a *relative  $\varepsilon_{\text{rel}}$ -close approximation*) of  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$  is a set

$$\mathcal{C} = \{(W, f_W, c_W) : \text{association rule } W, f_W \in [0, 1], c_W \in [0, 1]\}$$

of triplets  $(W, f_W, c_W)$  where  $f_W$  and  $c_W$  approximate  $f_{\mathcal{D}}(W)$  and  $c_{\mathcal{D}}(W)$  respectively.  $\mathcal{C}$  is such that:

1.  $\mathcal{C}$  contains all association rules appearing in  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ ;
2.  $\mathcal{C}$  contains no association rule  $W$  with frequency  $f_{\mathcal{D}}(W) < \theta - \varepsilon_{\text{abs}}$  (resp.  $f_{\mathcal{D}}(W) < (1 - \varepsilon_{\text{rel}})\theta$ );
3. For every triplet  $(W, f_W, c_W) \in \mathcal{C}$ , it holds  $|f_{\mathcal{D}}(W) - f_W| \leq \varepsilon_{\text{abs}}$  (resp.  $|f_{\mathcal{D}}(W) - f_W| \leq \varepsilon_{\text{rel}}\theta$ ).
4.  $\mathcal{C}$  contains no association rule  $W$  with confidence  $c_{\mathcal{D}}(W) < \gamma - \varepsilon_{\text{abs}}$  (resp.  $c_{\mathcal{D}}(W) < (1 - \varepsilon_{\text{rel}})\gamma$ );
5. For every triplet  $(W, f_W, c_W) \in \mathcal{C}$ , it holds  $|c_{\mathcal{D}}(W) - c_W| \leq \varepsilon_{\text{abs}}$  (resp.  $|c_{\mathcal{D}}(W) - c_W| \leq \varepsilon_{\text{rel}}c_{\mathcal{D}}(W)$ ).

Note that the definition of relative  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  (resp. to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ ) is more stringent than the definition of  $\varepsilon$ -close solution to frequent itemset mining (resp. association rule mining) in [6, Sect. 3]. Specifically, we require an approximation of the frequencies (and confidences) in addition to the approximation of the collection of itemsets or association rules (property 3 in Def. 4 and properties 3 and 5 in Def. 5).

## 2.2 VC-Dimension

The Vapnik-Chervonenkis (VC) Dimension of a space of points is a measure of the complexity or expressiveness of a family of indicator functions (or equivalently a family of subsets) defined on that space [36]. A finite bound on the VC-dimension of a structure implies a bound on the number of random samples required for approximately learning that structure. We outline here some basic definitions and results and refer the reader to the works of Alon and Spencer [3, Sect. 14.4], Chazelle [8, Chap. 4], and Vapnik [35] for more details on VC-dimension.

VC-dimension is defined on *range spaces*:

**Definition 6.** A *range space* is a pair  $(X, R)$  where  $X$  is a (finite or infinite) set and  $R$  is a (finite or infinite) family of subsets of  $X$ . The members of  $X$  are called *points* and those of  $R$  are called *ranges*.

To define the VC-dimension of a range space we consider the projection of the ranges into a set of points:

**Definition 7.** Let  $(X, R)$  be a range space and  $A \subset X$ . The *projection* of  $R$  on  $A$  is defined as  $P_R(A) = \{r \cap A : r \in R\}$ .

The definition of *shattered* set will be heavily used in our proofs:

**Definition 8.** Let  $(X, R)$  be a range space and  $A \subset X$ . If  $P_R(A) = 2^A$ , then  $A$  is said to be *shattered* by  $R$ .

The VC-dimension of a range space is the cardinality of the largest set shattered by the space:

**Definition 9.** Let  $S = (X, R)$  be a range space. The *Vapnik-Chervonenkis* dimension (or *VC-dimension*) of  $S$ , denoted as  $\text{VC}(S)$  is the maximum cardinality of a shattered subset of  $X$ . If there are arbitrary large shattered subsets, then  $\text{VC}(S) = \infty$ .

The main application of VC-dimension in statistics and learning theory is its relation to the size of the sample needed to approximate learning the ranges, in the following sense.

**Definition 10.** Let  $(X, R)$  be a range space and let  $A$  be a finite subset of  $X$ .

1. For  $0 < \varepsilon < 1$ , a subset  $B \subset A$  is an  $\varepsilon$ -approximation for  $A$  if  $\forall r \in R$ , we have

$$\left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon. \quad (1)$$

2. For  $0 < p, \varepsilon < 1$ , a subset  $B \subset A$  is a *relative*  $(p, \varepsilon)$ -approximation for  $A$  if for any range  $r \in R$  such that  $\frac{|A \cap r|}{|A|} \geq p$  we have

$$\left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \varepsilon \frac{|A \cap r|}{|A|} \text{ and for any range } r \in R \text{ such that } \frac{|A \cap r|}{|A|} < p \text{ we have } \frac{|B \cap r|}{|B|} \leq (1 + \varepsilon)p.$$

An  $\varepsilon$ -approximation (resp. a relative  $(p, \varepsilon)$ -approximation) can be constructed by random sampling points of the point space [16, Thm. 2.12 (resp. 2.11)] (see also [23]).

**Theorem 1.** *There is an absolute positive constant  $c$  (resp.  $c'$ ) such that if  $(X, R)$  is a range-space of VC-dimension at most  $d$ ,  $A \subset X$  is a finite subset and  $0 < \varepsilon, \delta < 1$  (resp. and  $0 < p < 1$ ), then a random subset  $B \subset A$  of cardinality  $m$ , where*

$$m \geq \min \left\{ |A|, \frac{c}{\varepsilon^2} \left( d + \log \frac{1}{\delta} \right) \right\}, \quad (2)$$

*(resp.  $m \geq \min \left\{ |A|, \frac{c'}{\varepsilon^2 p} \left( d \log \frac{1}{p} + \log \frac{1}{\delta} \right) \right\}$ ) is an  $\varepsilon$ -approximation (resp. a relative  $(p, \varepsilon)$ -approximation) for  $A$  with probability at least  $1 - \delta$ .*

Note that throughout the work we assume the sample to be drawn *with* replacement if  $m < |A|$  (otherwise the sample is exactly the set  $A$ ). Löffler and Phillips [25] showed experimentally that the absolute constant  $c$  is approximately 0.5. It is also interesting to note that an  $\varepsilon$ -approximation of size  $O(\frac{d}{\varepsilon^2} \log \frac{1}{\varepsilon})$  can be built *deterministically* in time  $O(d^{3d}(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon})^d |X|)$  [8].

### 3 The Dataset's Range Space and its VC-dimension

We define one range space that is used in the derivation of the sample sizes needed to approximate the solutions to the tasks of market basket analysis.

**Definition 11.** Let  $\mathcal{D}$  be a dataset of transactions that are subsets of a ground set  $\mathcal{I}$ . We define  $S = (X, R)$  to be a range space associated with  $\mathcal{D}$  such that:

1.  $X = \mathcal{D}$  is the set of transactions in the dataset.
2.  $R = \{T_{\mathcal{D}}(W) \mid W \subseteq \mathcal{I}\}$  is a family of sets of transactions such that for each itemset  $W \subseteq \mathcal{I}$ , the set  $T_{\mathcal{D}}(W) = \{\tau \in \mathcal{D} \mid W \subseteq \tau\}$  of all transactions containing  $W$  is an element of  $R$ .

**Theorem 2.** Let  $\mathcal{D}$  be a dataset and let  $S = (X, R)$  be the associated range space. Let  $d \in \mathbb{N}$ . Then  $\text{VC}(S) \geq d$  if and only if there exists a set  $\mathcal{A} \subseteq \mathcal{D}$  of  $d$  transactions from  $\mathcal{D}$  such that for each subset  $\mathcal{B} \subseteq \mathcal{A}$  of  $\mathcal{A}$ , there exists an itemset  $I_{\mathcal{B}}$  such that:

1. all transactions in  $\mathcal{B}$  contain  $I_{\mathcal{B}}$ , and
2. no transaction  $\rho \in \mathcal{A} \setminus \mathcal{B}$  contains  $I_{\mathcal{B}}$ .

*Proof.* “ $\Leftarrow$ ”. From the definition of  $I_{\mathcal{B}}$ , we have that  $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$ . By definition of  $P_R(\mathcal{A})$  this means that  $\mathcal{B} \in P_R(\mathcal{A})$ , for any subset  $\mathcal{B}$  of  $\mathcal{A}$ . Then  $P_R(\mathcal{A}) = 2^{\mathcal{A}}$ , which implies  $\text{VC}(S) \geq d$ .

“ $\Rightarrow$ ”. Let  $\text{VC}(S) \geq d$ . Then by definition of VC-Dimension there is a set  $\mathcal{A} \subseteq \mathcal{D}$  of  $d$  transactions from  $\mathcal{D}$  such that  $P_R(\mathcal{A}) = 2^{\mathcal{A}}$ . By definition of  $P_R(\mathcal{A})$ , this means that for each subset  $\mathcal{B} \subseteq \mathcal{A}$  there exists an itemset  $I_{\mathcal{B}}$  such that  $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$ . We want to show that no transaction  $\rho \in \mathcal{A} \setminus \mathcal{B}$  contains  $I_{\mathcal{B}}$ . Assume now by contradiction that there is a transaction  $\rho^* \in \mathcal{A} \setminus \mathcal{B}$  containing  $I_{\mathcal{B}}$ . Then  $\rho^* \in T_{\mathcal{D}}(I_{\mathcal{B}})$  and, given that  $\rho^* \in \mathcal{A}$ , we have  $\rho^* \in T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A}$ . But by construction, we have that  $T_{\mathcal{D}}(I_{\mathcal{B}}) \cap \mathcal{A} = \mathcal{B}$  and  $\rho^* \notin \mathcal{B}$  because  $\rho^* \in \mathcal{A} \setminus \mathcal{B}$ . Then we have a contradiction, and there can not be such a transaction  $\rho^*$ .  $\square$

**Corollary 1.** Let  $\mathcal{D}$  be a dataset and  $S = (\mathcal{D}, R)$  be the corresponding range space. Then, the VC-Dimension  $\text{VC}(S)$  of  $S$ , is the maximum integer  $d$  such that there is a set  $\mathcal{A} \subseteq \mathcal{D}$  of  $d$  transactions from  $\mathcal{D}$  such that for each subset  $\mathcal{B} \subseteq \mathcal{A}$  of  $\mathcal{A}$ , there exists an itemset  $I_{\mathcal{B}}$  such that

1. all transactions in  $\mathcal{B}$  contain  $I_{\mathcal{B}}$ , and
2. no transaction  $\rho \in \mathcal{A} \setminus \mathcal{B}$  contains  $I_{\mathcal{B}}$ .

Computing the exact VC-dimension of a dataset is extremely expensive from a computational point of view. This does not come as a surprise, as it is known that computing the VC-dimension of a range space  $(X, R)$  can take time  $O(|R||X|^{\log |R|})$  [24, Thm. 4.1]. It is instead possible to give an upper bound to the VC-dimension of a dataset, and a procedure to efficiently compute the bound.

**Definition 12.** Let  $\mathcal{D}$  be a dataset. The  $d$ -index of a dataset is defined as the maximum integer  $d$  such that  $\mathcal{D}$  contains at least  $d$  transactions of length at least  $d$ .

A note of folklore: if the dataset represents the scientific publications of a given scientist, with transactions corresponding to articles and items in a transaction corresponding to the citations received by the paper, then the  $d$ -index of the dataset corresponds to the  $h$ -index of the scientist.

The  $d$ -index  $d$  of a dataset  $\mathcal{D}$  can be computed in one scan of the dataset and with total memory  $O(d)$ . The scan starts with  $d^* = 1$  and it stores the length of the first transaction. At any given step the procedure stores  $d^*$ , the current estimate of  $d$ , computed as the maximum  $d'$  such that the scan up to this step found at least  $d'$  transactions with length at least  $d'$ , and keeps a list of the sizes of the transactions longer than  $d'$  found so far. There can be no more than  $d'$  such transactions. As the scan proceeds, the procedure updates  $d^*$  and the list of transactions sizes greater than  $d^*$ .

The  $d$ -index is an upper bound to the VC-dimension of a dataset.

**Theorem 3.** Let  $\mathcal{D}$  be a dataset with  $d$ -index  $d$ . Then the range space  $S = (X, R)$  corresponding to  $\mathcal{D}$  has VC-dimension at most  $d$ .

*Proof.* Let  $\ell > d$  and assume that  $S$  has VC-dimension  $\ell$ . From Def. 10 there is a set  $\mathcal{K}$  of  $\ell$  transactions that is shattered by  $R$ . By definition of  $d$  and  $\ell$ ,  $\mathcal{K}$  must contain a transaction  $\tau$  such that  $|\tau| \leq d$ . The transaction  $\tau$  is a member of  $2^{\ell-1}$  subsets of  $\mathcal{K}$ . We denote these subsets of  $\mathcal{K}$  containing  $\tau$  as  $\mathcal{A}_{\tau}^{(i)}$ ,  $1 \leq i \leq 2^{\ell-1}$ , labeling them in an arbitrary order. Since  $\mathcal{K}$  is shattered (i.e.,  $P_R(\mathcal{K}) = 2^{\mathcal{K}}$ ), we have

$$\mathcal{A}_{\tau}^{(i)} \in P_R(\mathcal{K}), 1 \leq i \leq 2^{\ell-1}.$$

From the above and the definition of  $P_R(\mathcal{K})$ , it follows that for each set of transactions  $\mathcal{A}_{\tau}^{(i)}$  there must be a non-empty itemset  $B_{\tau}^{(i)}$  such that

$$T_{\mathcal{D}}(B_{\tau}^{(i)}) \cap \mathcal{K} = \mathcal{A}_{\tau}^{(i)} \in P_R(\mathcal{K}). \quad (3)$$

Since the  $\mathcal{A}_\tau^{(i)}$  are all different from each other, this means that the  $T_{\mathcal{D}}(B_\tau^{(i)})$  are all different from each other, which in turn requires that the  $B_\tau^{(i)}$  be all different from each other, for  $1 \leq i \leq 2^{\ell-1}$ .

Since  $\tau \in \mathcal{A}_\tau^{(i)}$  and  $\tau \in \mathcal{K}$  by construction, it follows from (3) that

$$\tau \in T_{\mathcal{D}}(B_\tau^{(i)}), 1 \leq i \leq 2^{\ell-1}.$$

From the above and the definition of  $T_{\mathcal{D}}(B_\tau^{(i)})$ , we get that all the itemsets  $B_\tau^{(i)}, 1 \leq i \leq 2^{\ell-1}$  appear in the transaction  $\tau$ . But  $|\tau| \leq d < \ell$ , therefore  $\tau$  can only contain at most  $2^d - 1 < 2^{\ell-1}$  non-empty itemsets, while there are  $2^{\ell-1}$  different itemsets  $B_\tau^{(i)}$ .

This is a contradiction, therefore our assumption is false and  $\mathcal{K}$  cannot be shattered by  $R$ , which implies that  $\text{VC}(S) \leq d$ .  $\square$

This bound is strict, i.e., there are indeed datasets with VC-dimension exactly  $d$ , as formalized by the following Theorem.

**Theorem 4.** *There exists a dataset  $\mathcal{D}$  with  $d$ -index  $d$  and such the corresponding range space has VC-dimension exactly  $d$ .*

*Proof.* For  $d = 1$ ,  $\mathcal{D}$  can be any dataset with transactions of length 1. Let  $\tau$  be any transaction in  $\mathcal{D}$  and let  $a$  be the item in  $\tau$ . The set  $\{\tau\} \subseteq \mathcal{D}$  is shattered because  $T_{\mathcal{D}}(a) \cap \{\tau\} = \{\tau\}$  and  $\emptyset \cap \{\tau\} = \emptyset$ .

Without loss of generality, let the ground set  $\mathcal{I}$  be  $\mathbb{N}$ . For a fixed  $d > 1$ , let  $\tau_i = \{0, 1, 2, \dots, i-1, i+1, \dots, d\}$   $1 \leq i \leq d$ , and consider the set of  $d$  transactions  $\mathcal{K} = \{\tau_i, 1 \leq i \leq d\}$ . Note that  $|\tau_i| = d$  and  $|\mathcal{K}| = d$ .

$\mathcal{D}$  is a dataset containing  $\mathcal{K}$  and any number of arbitrary transactions from  $2^{\mathcal{I}}$  of length at most  $d$ . Let  $S = (X, R)$  be the range space corresponding to  $\mathcal{D}$ . We now show that  $\mathcal{K} \subseteq X$  is shattered by ranges from  $R$ , which implies  $\text{VC}(S) \geq d$ .

For each  $\mathcal{A} \in 2^{\mathcal{K}} \setminus \{\mathcal{K}, \emptyset\}$ , let  $Y_{\mathcal{A}}$  be the itemset

$$Y_{\mathcal{A}} = \{1, \dots, d\} \setminus \{i : \tau_i \in \mathcal{A}\}.$$

Let  $Y_{\mathcal{K}} = \{0\}$  and let  $Y_{\emptyset} = \{d+1\}$ . By construction we have

$$T_{\mathcal{K}}(Y_{\mathcal{A}}) = \mathcal{A}, \forall \mathcal{A} \in 2^{\mathcal{K}}$$

i.e., the itemset  $Y_{\mathcal{A}}$  appears in all transactions in  $\mathcal{A} \subseteq \mathcal{K}$  but not in any transaction from  $\mathcal{K} \setminus \mathcal{A}$ ,  $\forall \mathcal{A} \in 2^{\mathcal{K}}$ . This means that

$$T_{\mathcal{D}}(Y_{\mathcal{A}}) \cap \mathcal{K} = T_{\mathcal{K}}(Y_{\mathcal{A}}) = \mathcal{A}, \forall \mathcal{A} \in 2^{\mathcal{K}}.$$

Since  $\forall \mathcal{A} \in 2^{\mathcal{K}}, T_{\mathcal{D}}(Y_{\mathcal{A}}) \in R$  by construction, the above implies that

$$\mathcal{A} \in P_R(\mathcal{K}), \forall \mathcal{A} \in 2^{\mathcal{K}}$$

This means that  $\mathcal{K}$  is shattered by  $R$ , hence  $\text{VC}(S) \geq d$ . From this and Thm. 3, we can conclude that  $\text{VC}(S) = d$ .  $\square$

The datasets built in the proof of Thm. 4 are extremely artificial. Our experiments suggest that the VC-dimension of real datasets is usually much smaller than the upper bound presented in Thm. 3.

## 4 Mining (top- $K$ ) Frequent Itemsets and Association Rules

We apply the VC-dimension results to constructing efficient sampling algorithms with performance guarantees for approximating the collections of FI's, top- $K$  FI's and AR's.

### 4.1 Mining Frequent Itemsets

We construct bounds for the sample size needed to obtain relative/absolute  $\varepsilon$ -close approximations to the collection of FI's. The algorithms to compute the approximations use a standard exact FI's mining algorithm on the sample, with an appropriately adjusted minimum frequency threshold, as formalized in the following lemma.

**Lemma 1.** *Let  $\mathcal{D}$  be a dataset with transactions built on a ground set  $\mathcal{I}$ , and let  $d$  be the  $d$ -index of  $\mathcal{D}$ . Let  $0 < \varepsilon, \delta < 1$ . Let  $S$  be a random sample of  $\mathcal{D}$  with size  $|\mathcal{S}| = \min\{|\mathcal{D}|, \frac{4c}{\varepsilon^2} (d + \log \frac{1}{\delta})\}$ , for some absolute constant  $c$ . Then  $\text{FI}(\mathcal{S}, \mathcal{I}, \theta - \frac{\varepsilon}{2})$  is an absolute  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  with probability at least  $1 - \delta$ .*

*Proof.* Suppose that  $S$  is a  $\frac{\varepsilon}{2}$ -approximation of the range space  $(X, R)$  corresponding to  $\mathcal{D}$ . From Thm. 1 we know that this happens with probability at least  $1 - \delta$ . This means that  $\forall X \in 2^{\mathcal{I}}, f_{\mathcal{S}}(X) \in [f_{\mathcal{D}}(X) - \frac{\varepsilon}{2}, f_{\mathcal{D}}(X) + \frac{\varepsilon}{2}]$ . This holds in particular for the itemsets in  $\mathcal{C} = \text{FI}(\mathcal{S}, \mathcal{I}, \theta - \frac{\varepsilon}{2})$ , which therefore satisfies property 3 from Def. 4. It also means that  $\forall X \in \text{FI}(\mathcal{D}, \mathcal{I}, \theta), f_{\mathcal{S}}(X) \geq \theta - \frac{\varepsilon}{2}$ , so  $\mathcal{C}$  also guarantees property 1 from Def. 4. Let now  $Y \in 2^{\mathcal{I}}$  be such that  $f_{\mathcal{D}}(Y) < \theta - \varepsilon$ . Then, for the properties of  $S$ ,  $f_{\mathcal{S}}(Y) < \theta - \frac{\varepsilon}{2}$ , i.e.,  $Y \notin \mathcal{C}$ , which allows us to conclude that  $\mathcal{C}$  also has property 2 from Def. 4.  $\square$

One very interesting consequence of this result is that we do not need to know the minimum frequency threshold  $\theta$  in advance to build the sample: the properties of the  $\varepsilon$ -approximation allow to use the same sample for any threshold and for different thresholds, i.e., the sample does not need to be rebuilt if we want to mine it with a threshold  $\theta$  first and with another threshold  $\theta'$  later.

It is important to note that the VC-dimension of a dataset, and therefore the sample size from (2) needed to probabilistically obtain an  $\varepsilon$ -approximation, is independent from the size (number of transactions) in  $\mathcal{D}$  and also of the size of  $\text{FI}(\mathcal{S}, \mathcal{I}, \theta)$ . It only depends on the quantity  $d$ , which is always less or equal to the length of the longest transaction in the dataset, which in turn is less or equal to the number of different items  $|\mathcal{I}|$ .

To obtain a relative  $\varepsilon$ -close approximation, we need to add a dependency on  $\theta$  as shown in the following Lemma.

**Lemma 2.** *Let  $\mathcal{D}$ ,  $d$ ,  $\varepsilon$ , and  $\delta$  as in Lemma 1. Let  $\mathcal{S}$  be a random sample of  $\mathcal{D}$  with size*

$$|\mathcal{S}| = \min\{|\mathcal{D}|, \frac{4(2+\varepsilon)c}{\varepsilon^2\theta(2-\varepsilon)} \left( d \log \frac{2+\varepsilon}{\theta(2-\varepsilon)} + \log \frac{1}{\delta} \right) \},$$

*for some absolute constant  $c$ . Then  $\text{FI}(\mathcal{S}, \mathcal{I}, (1 - \frac{\varepsilon}{2})\theta)$  is a relative  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  with probability at least  $1 - \delta$ .*

*Proof.* Let  $p = \theta \frac{2-\varepsilon}{2+\varepsilon}$ . From Thm. 1, the sample  $\mathcal{S}$  is a relative  $(p, \varepsilon/2)$ -approximation of the range space associated to  $\mathcal{D}$  with probability at least  $1 - \delta$ . For any itemset  $X$  in  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ , we have  $f_{\mathcal{D}}(X) \geq \theta > p$ , so  $f_{\mathcal{S}}(X) \geq (1 - \varepsilon/2)f_{\mathcal{D}}(X) \geq (1 - \varepsilon/2)\theta$ , which implies  $X \in \text{FI}(\mathcal{S}, \mathcal{I}, (1 - \frac{\varepsilon}{2})\theta)$ , so property 1 from Def. 4 holds. Let now  $X$  be an itemsets with  $f_{\mathcal{D}}(X) < (1 - \varepsilon)\theta$ . From our choice of  $p$ , we always have  $p > (1 - \varepsilon)\theta$ , so  $f_{\mathcal{S}}(X) \leq p(1 + \varepsilon/2) < \theta(1 - \varepsilon/2)$ . This means  $X \notin \text{FI}(\mathcal{S}, \mathcal{I}, (1 - \frac{\varepsilon}{2})\theta)$ , as requested by property 2 from Def. 4. Since  $(1 - \frac{\varepsilon}{2})\theta = p(1 + \frac{\varepsilon}{2})$ , it follows that only itemsets  $X$  with  $f_{\mathcal{D}}(X) \geq p$  can be in  $\text{FI}(\mathcal{S}, \mathcal{I}, (1 - \frac{\varepsilon}{2})\theta)$ . For these itemsets it holds  $|f_{\mathcal{S}}(X) - f_{\mathcal{D}}(X)| \leq \frac{\varepsilon}{2}f_{\mathcal{D}}(X)$ , as requested by property 3 from Def. 4.  $\square$

## 4.2 Mining Top- $K$ Frequent Itemsets

Given the equivalence  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K) = \text{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)})$ , we could use the above FI's sampling algorithms if we had a good approximation of  $f_{\mathcal{D}}^{(K)}$ , the threshold frequency of the top- $K$  FI's.

For the absolute  $\varepsilon$ -close approximation we first execute a standard top- $K$  FI's mining algorithm on the sample to estimate  $f_{\mathcal{D}}^{(K)}$  and then run a standard FI's mining algorithm on the same sample using a minimum frequency threshold depending on our estimate of  $f_{\mathcal{S}}^{(K)}$ . Lemma 3 formalizes this intuition.

**Lemma 3.** *Let  $\mathcal{D}$ ,  $d$ ,  $\varepsilon$ , and  $\delta$  be as in Lemma 1. Let  $K$  be a positive integer. Let  $\mathcal{S}$  be a random sample of  $\mathcal{D}$  with size  $|\mathcal{S}| = \min\{|\mathcal{D}|, \frac{16c}{\varepsilon^2} (d + \log \frac{1}{\delta})\}$ , for some absolute constant  $c$ , then  $\text{FI}(\mathcal{S}, \mathcal{I}, f_{\mathcal{S}}^{(K)} - \frac{\varepsilon}{2})$  is an absolute  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  with probability at least  $1 - \delta$ .*

*Proof.* Suppose that  $\mathcal{S}$  is a  $\frac{\varepsilon}{4}$ -approximation of the range space  $(X, R)$  corresponding to  $\mathcal{D}$ . From Thm. 1 we know that this happens with probability at least  $1 - \delta$ . This means that,  $\forall Y \in 2^{\mathcal{I}}, f_{\mathcal{S}}(Y) \in [f_{\mathcal{D}}(Y) - \frac{\varepsilon}{4}, f_{\mathcal{D}}(Y) + \frac{\varepsilon}{4}]$ . Consider now  $f_{\mathcal{S}}^{(K)}$ , the frequency of the  $K$ -th most frequent itemset in the sample. Clearly,  $f_{\mathcal{S}}^{(K)} \geq f_{\mathcal{D}}^{(K)} - \frac{\varepsilon}{4}$ , because there are at least  $K$  itemsets (for example any subset of size  $K$  of  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ ) with frequency in the sample at least  $f_{\mathcal{D}}^{(K)} - \frac{\varepsilon}{4}$ . On the other hand  $f_{\mathcal{S}}^{(K)} \leq f_{\mathcal{D}}^{(K)} + \frac{\varepsilon}{4}$ , because there cannot be  $K$  itemsets with a frequency in the sample greater than  $f_{\mathcal{D}}^{(K)} + \frac{\varepsilon}{4}$ : only itemsets with frequency in the dataset strictly greater than  $f_{\mathcal{D}}^{(K)}$  can have a frequency in the sample greater than  $f_{\mathcal{D}}^{(K)} + \frac{\varepsilon}{4}$ , and there are at most  $K - 1$  such itemsets. Let now  $\eta = f_{\mathcal{S}}^{(K)} - \frac{\varepsilon}{2}$ , and consider  $\text{FI}(\mathcal{S}, \mathcal{I}, \eta)$ . We have  $\eta \leq f_{\mathcal{D}}^{(K)} - \frac{\varepsilon}{4}$ , so for the properties of  $\mathcal{S}$ ,  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K) = \text{FI}(\mathcal{D}, \mathcal{I}, f_{\mathcal{D}}^{(K)}) \subseteq \text{FI}(\mathcal{S}, \mathcal{I}, \eta)$ , which then guarantees property 1 from Def. 4. On the other hand, let  $Y$  be an itemset such that  $f_{\mathcal{D}}(Y) < f_{\mathcal{D}}^{(K)} - \varepsilon$ . Then  $f_{\mathcal{S}}(Y) < f_{\mathcal{D}}^{(K)} - \frac{3}{4}\varepsilon \leq \eta$ , so  $Y \notin \text{FI}(\mathcal{S}, \mathcal{I}, \eta)$ , corresponding to property 2 from Def. 4. Property 3 from Def. 4 follows from the properties of  $\mathcal{S}$ .  $\square$

Note that as in the case of the sample size required for an absolute  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ , we do not need to know  $K$  in advance to compute the sample size for obtaining an absolute  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ .

Two different samples are needed for computing a relative  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ , the first one to compute a lower bound to  $f_{\mathcal{D}}^{(K)}$ , the second to extract the approximation. Details for this case are presented in Lemma 4.

**Lemma 4.** *Let  $\mathcal{D}$ ,  $d$ ,  $\varepsilon$ , and  $\delta$  be as in Lemma 1. Let  $K$  be a positive integer. Let  $\delta_1, \delta_2$  be two reals such that  $(1 - \delta_1)(1 - \delta_2) \geq (1 - \delta)$ . Let  $\mathcal{S}_1$  be a random sample of  $\mathcal{D}$  with some size  $|\mathcal{S}_1| = \frac{\phi c}{\varepsilon^2} (d + \log \frac{1}{\delta_1})$  for some  $\phi > 2\sqrt{2}/\varepsilon$  and some absolute constant  $c$ . If  $f_{\mathcal{S}_1}^{(K)} \geq \frac{2\sqrt{2}}{\varepsilon\phi}$ , then let  $p = \frac{2-\varepsilon}{2+\varepsilon}\theta$  and let  $\mathcal{S}_2$  be a random sample of  $\mathcal{D}$  of size  $\min\{|\mathcal{D}|, \frac{4c}{\varepsilon^2 p} (d \log \frac{1}{p} + \log \frac{1}{\delta})\}$  for some absolute constant  $c$ . Then  $\text{FI}(\mathcal{S}_2, \mathcal{I}, (1 - \varepsilon/2)(f_{\mathcal{S}_1}^{(K)} - \varepsilon/\sqrt{2\phi}))$  is a relative  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$  with probability at least  $1 - \delta$ .*



*Proof.* Assume that  $\mathcal{S}_1$  is a  $\frac{\varepsilon}{\sqrt{2\phi}}$ -approximation for  $\mathcal{D}$  and  $\mathcal{S}_2$  is a relative  $(p, \varepsilon/2)$ -approximation for  $\mathcal{D}$ . The probability of these two events happening at the same time is at least  $1 - \delta$ , from Thm. 1.

Following the steps of the proof of Lemma 3 we can easily get that, from the properties of  $\mathcal{S}_1$ ,

$$f_{\mathcal{S}_1}^{(K)} - \frac{\varepsilon}{\sqrt{2\phi}} \leq f_{\mathcal{D}}^{(K)} \leq f_{\mathcal{S}_1}^{(K)} + \frac{\varepsilon}{\sqrt{2\phi}}. \quad (4)$$

Consider now an element  $X \in \text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ . We have by definition  $f_{\mathcal{D}}(X) \geq f_{\mathcal{D}}^{(K)} > f_{\mathcal{S}_1}^{(K)} - \frac{\varepsilon}{\sqrt{2\phi}} \geq p$ , and from the properties of  $\mathcal{S}_2$ , it follows that  $f_{\mathcal{S}}(X) \geq (1 - \varepsilon/2)f_{\mathcal{D}}(X) \geq (1 - \varepsilon/2)(f_{\mathcal{S}_1}^{(K)} - \frac{\varepsilon}{\sqrt{2\phi}})$ , which implies  $X \in \text{FI}(\mathcal{S}_2, \mathcal{I}, (1 - \varepsilon/2)(f_{\mathcal{S}_1}^{(K)} - \varepsilon/\sqrt{2\phi}))$  and therefore property 1 from Def. 4 holds for  $\text{FI}(\mathcal{S}_2, \mathcal{I}, \eta)$ .

Let now  $Y$  be an itemset such that  $f_{\mathcal{D}}(Y) < (1 - \varepsilon)f_{\mathcal{D}}^{(K)}$ . From our choice of  $p$  we have that  $f_{\mathcal{D}}(A) < p$ . Then  $f_{\mathcal{S}_2}(A) < (1 + \varepsilon/2)p < (1 - \varepsilon/2)(f_{\mathcal{S}_1}^{(K)} - \frac{\varepsilon}{\sqrt{2\phi}})$ . Therefore,  $Y \notin \text{FI}(\mathcal{S}_2, \mathcal{I}, \eta)$  and property 2 from Def. 4 is guaranteed.

Property 3 from Def. 4 follows from (4) and the properties of  $\mathcal{S}_2$ .  $\square$

### 4.3 Mining Association Rules

Our final theoretical contribution concerns the discovery of relative/absolute approximations to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \eta)$  from a sample. Lemma 5 builds on a result from [6, Sect. 5] and covers the *relative* case, while Lemma 6 deals with the *absolute* one.

**Lemma 5.** Let  $0 < \delta, \varepsilon, \theta, \gamma < 1$ ,  $\phi = \max\{3, 2 - \varepsilon + 2\sqrt{1 - \varepsilon}\}$ ,  $\eta = \frac{\varepsilon}{\phi}$ , and  $p = \frac{1 - \eta}{1 + \eta}\theta$ . Let  $\mathcal{D}$  be a dataset with  $d$ -index  $d$ . Let  $\mathcal{S}$  be a random sample of  $\mathcal{D}$  of size  $\min\{|\mathcal{D}|, \frac{c}{\eta^2 p}(d \log \frac{1}{p} + \log \frac{1}{\delta})\}$  for some absolute constant  $c$ . Then  $\text{AR}(\mathcal{S}, \mathcal{I}, (1 - \eta)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$  is a relative  $\varepsilon$ -close approximation to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$  with probability at least  $1 - \delta$ .

*Proof.* Suppose  $\mathcal{S}$  is a relative  $(p, \eta)$ -approximation for the range space corresponding to  $\mathcal{D}$ . From Thm. 1 we know this happens with probability at least  $1 - \delta$ .

Let  $W \in \text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$  be the association rule “ $A \Rightarrow B$ ”, where  $A$  and  $B$  are itemsets. By definition  $f_{\mathcal{D}}(W) = f_{\mathcal{D}}(A \cup B) \geq \theta > p$ . From this and the properties of  $\mathcal{S}$ , we get

$$f_{\mathcal{S}}(W) = f_{\mathcal{S}}(A \cup B) \geq (1 - \eta)f_{\mathcal{D}}(A \cup B) \geq (1 - \eta)\theta.$$

Note that, from the fact that  $f_{\mathcal{D}}(W) = f_{\mathcal{D}}(A \cup B) \geq \theta$ , it follows that  $f_{\mathcal{D}}(A), f_{\mathcal{D}}(B) \geq \theta > p$ , for the anti-monotonicity property of the frequency of itemsets.

By definition,  $c_{\mathcal{D}}(W) = \frac{f_{\mathcal{D}}(W)}{f_{\mathcal{D}}(A)} \geq \gamma$ . Then,

$$c_{\mathcal{S}}(W) = \frac{f_{\mathcal{S}}(W)}{f_{\mathcal{S}}(A)} \geq \frac{(1 - \eta)f_{\mathcal{D}}(W)}{(1 + \eta)f_{\mathcal{D}}(A)} \geq \frac{1 - \eta}{1 + \eta} \cdot \frac{f_{\mathcal{D}}(W)}{f_{\mathcal{D}}(A)} \geq \frac{1 - \eta}{1 + \eta}\gamma.$$

It follows that  $W \in \text{AR}(\mathcal{S}, \mathcal{I}, (1 - \eta)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$ , hence property 1 from Def. 5 is satisfied.

Let now  $Z$  be the association rule “ $C \Rightarrow D$ ”, such that  $f_{\mathcal{D}}(Z) = f_{\mathcal{D}}(C \cup D) < (1 - \varepsilon)\theta$ . But from our definitions of  $\eta$  and  $p$ , it follows that  $f_{\mathcal{D}}(Z) < p < \theta$ , hence  $f_{\mathcal{S}}(Z) < (1 + \eta)p < (1 - \eta)\theta$ , and therefore  $Z \notin \text{AR}(\mathcal{S}, \mathcal{I}, (1 - \eta)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$ , as requested by property 2 from Def. 5.

Consider now an association rule  $Y = “E \Rightarrow F”$  such that  $c_{\mathcal{D}}(Y) < (1 - \varepsilon)\gamma$ . Clearly, we are only concerned with  $Y$  such that  $f_{\mathcal{D}}(Y) \geq p$ , otherwise we just showed that  $Y$  can not be in  $\text{AR}(\mathcal{S}, \mathcal{I}, (1 - \eta)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$ . From this and the anti-monotonicity property, it follows that  $f_{\mathcal{D}}(E), f_{\mathcal{D}}(F) \geq p$ . Then,

$$c_{\mathcal{S}}(Y) = \frac{f_{\mathcal{S}}(Y)}{f_{\mathcal{S}}(E)} \leq \frac{(1 + \eta)f_{\mathcal{D}}(Y)}{(1 - \eta)f_{\mathcal{D}}(E)} < \frac{1 + \eta}{1 - \eta}(1 - \varepsilon)\gamma < \frac{1 - \eta}{1 + \eta}\gamma,$$

where the last inequality follows from the fact that  $(1 - \eta)^2 > (1 + \eta)(1 - \varepsilon)$  for our choice of  $\eta$ . We can conclude that  $Y \notin \text{AR}(\mathcal{S}, \mathcal{I}, (1 - \varepsilon)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$  and therefore property 4 from Def. 5 holds.

Properties 3 and 5 from Def. 5 follow from the above steps (i.e., what association rules can be in the approximations), from the definition of  $\phi$ , and from the properties of  $\mathcal{S}$ .  $\square$

**Lemma 6.** Let  $\mathcal{D}$ ,  $d$ ,  $\theta$ ,  $\gamma$ ,  $\varepsilon$ , and  $\delta$  be as in Lemma 5 and let  $\varepsilon_{\text{rel}} = \frac{\varepsilon}{\max\{\theta, \gamma\}}$ .

Fix  $\phi = \max\{3, 2 - \varepsilon_{\text{rel}} + 2\sqrt{1 - \varepsilon_{\text{rel}}}\}$ ,  $\eta = \frac{\varepsilon_{\text{rel}}}{\phi}$ , and  $p = \frac{1 - \eta}{1 + \eta}\theta$ . Let  $\mathcal{S}$  be a random sample of  $\mathcal{D}$  of size  $\min\{|\mathcal{D}|, \frac{c}{\eta^2 p}(d \log \frac{1}{p} + \log \frac{1}{\delta})\}$  for some absolute constant  $c$ . Then  $\text{AR}(\mathcal{S}, \mathcal{I}, (1 - \eta)\theta, \frac{1 - \eta}{1 + \eta}\gamma)$  is an absolute  $\varepsilon$ -close approximation to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ .

*Proof.* The thesis follows from Lemma 5 by setting  $\varepsilon$  there to  $\varepsilon_{\text{rel}}$ .  $\square$

Note that the sample size needed for absolute  $\varepsilon$ -close approximations to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$  depends on  $\theta$  and  $\gamma$ , which was not the case for absolute  $\varepsilon$ -close approximations to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  and  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ .

## 5 Experimental Evaluation

In this section we present an extensive experimental evaluation of our methods to extract approximations of  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ ,  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ , and  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ . Due to space constraints, we focus on a subset of the results.

Our first goal is to evaluate the *quality* of the approximations obtained using our methods, by comparing the experimental results to the analytical bounds. We also evaluate how strict the bounds are by testing whether the same quality of results can be achieved at sample sizes smaller than those suggested by the theoretical analysis. We then show that our methods can significantly speed-up the mining process, fulfilling the motivating promises of the use of sampling in the market basket analysis tasks. Lastly, we compare the sample sizes from our results to the best previous work [6].

We tested our methods on both real and artificial datasets. The real datasets come from the FIMI'04 repository (<http://fimi.ua.ac.be/data/>). Since most of them have a moderately small size, we replicated their transactions a number of times, with the only effect of increasing the size of the dataset but no change in the distribution of the frequencies of the itemsets. The artificial datasets were built such that their corresponding range spaces have VC-dimension equal to the maximum transaction length  $d$ , which is the maximum possible as shown in Thm. 3. To create these datasets, we followed the proof of Thm. 4 and used the generator included in ARtool (<http://www.cs.umb.edu/~laur/ARtool/>), which is similar to the one presented in [2]. We used the FP-Growth and Apriori implementations in ARtool to extract frequent itemsets and association rules. In all our experiments we fixed  $\delta = 0.1$ . In the experiments involving absolute (resp. relative)  $\varepsilon$ -close approximations we set  $\varepsilon = 0.01$  (resp.  $\varepsilon = 0.05$ ). The absolute constant  $c$  was fixed to 0.5 as suggested by [25]. For each dataset we selected a range of minimum frequency thresholds and a set of values for  $K$  when extracting the top- $K$  frequent itemsets. For association rules discovery we set the minimum confidence threshold  $\gamma \in \{0.5, 0.75, 0.9\}$ . For each dataset and each combination of parameters we created random samples with size as suggested by our theorems and with smaller sizes to evaluate the strictness of the bounds. We measured, for each set of parameters, the *absolute frequency error* and the *absolute confidence error*, defined as the error  $|f_{\mathcal{D}}(X) - f_{\mathcal{S}}(X)|$  (resp.  $|c_{\mathcal{D}}(Y) - c_{\mathcal{S}}(Y)|$ ) for an itemset  $X$  (resp. an association rule  $Y$ ) in the approximate collection extracted from sample  $\mathcal{S}$ . When dealing with the problem of extracting *relative*  $\varepsilon$ -close approximations, we defined the *relative frequency error* to be the absolute frequency error divided by the real frequency of the itemset and analogously for the relative confidence error (dividing by the real confidence). In the figures we plot the maximum and the average for these quantities, taken over all itemsets or association rules in the output collection. In order to limit the influence of a single sample, we computed and plot in the figures the maximum (resp. the average) of these quantities in three runs of our methods on three different samples for each size.

The first important result of our experiments is that, for all problems, for every combination of parameters and every run, the collection of itemsets or of association rules obtained using our methods always satisfied the requirements to be an absolute or relative  $\varepsilon$ -close approximation to the real collection. Thus in practice our methods indeed achieve or exceed the theoretical guarantees for approximations of the collections  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ ,  $\text{TOPK}(\mathcal{D}, \mathcal{I}, \theta)$ , and  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$ .

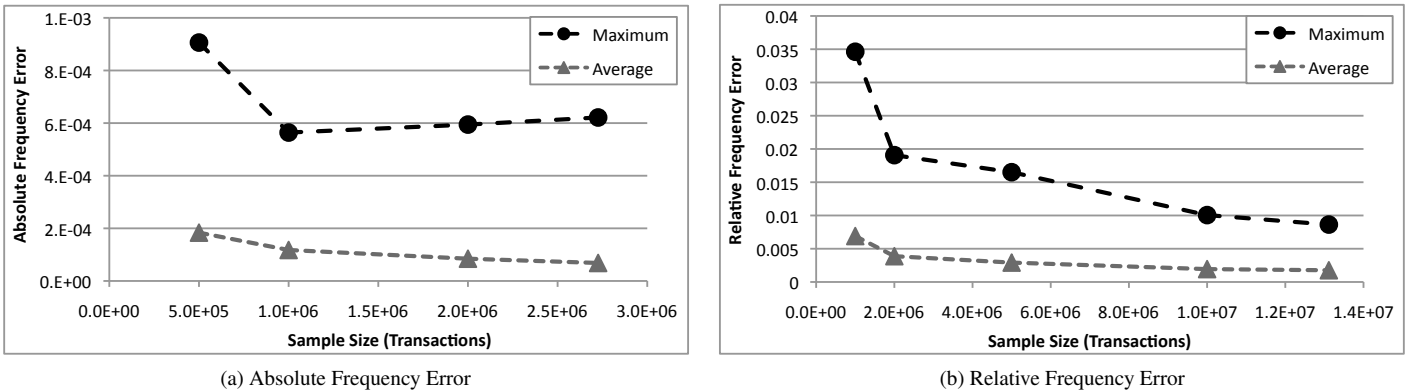


Figure 1: Absolute / Relative  $\varepsilon$ -close Approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$

Evaluating the strictness of the bounds to the sample size was the second goal of our experiments. In Figure 1a we show the behaviour of the maximum frequency error as function of the sample size in the itemsets obtained from samples using the method presented in Lemma 1 (i.e., we are looking for an *absolute*  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ ). The rightmost plotted point corresponds to the sample size suggested by the theoretical analysis. We are showing the results for the dataset BMS-POS replicated 40 times (d-index  $d = 134$ ), mined with  $\theta = 0.02$ . It is clear from the picture that the guaranteed error bounds are achieved even at sample sizes smaller than what suggested by the analysis and that the error at the sample size derived from the theory (rightmost plotted point for each line) is one to two orders of magnitude smaller than the maximum tolerable error  $\varepsilon = 0.01$ . This fact seems to suggest that there is still room for improvement in the bounds to the sample size needed to achieve an absolute  $\varepsilon$ -close approximation

to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ . In Fig. 1b we report similar results for the problem of computing a *relative*  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  for an artificial dataset whose range space has VC-dimension  $d$  equal to the length of the longest transaction in the dataset, in this case 33. The dataset contained 100 million transactions. The sample size, suggested by Lemma 2, was computed using  $\theta = 0.01$ ,  $\varepsilon = 0.05$ , and  $\delta = 0.1$ . The conclusions we can draw from the results for the behaviour of the relative frequency error are similar to those we got for the absolute case. For the case of absolute and relative  $\varepsilon$ -close approximation to  $\text{TOPK}(\mathcal{D}, \mathcal{I}, K)$ , we observed similar results, which we do not report here because of space constraints.

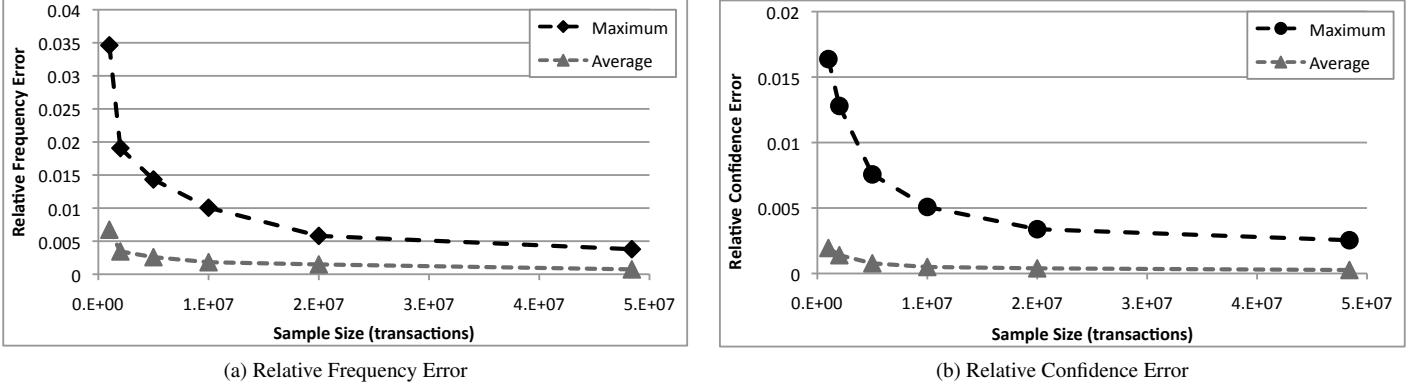


Figure 2: Relative  $\varepsilon$ -close approximation to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$

The results of the experiments to evaluate our method to extract a relative  $\varepsilon$ -close approximation to  $\text{AR}(\mathcal{D}, \mathcal{I}, \theta, \gamma)$  are presented in Fig. 2a and 2b. The same observations as before hold for the relative frequency error, while it is interesting to note that the relative confidence error is even smaller than the frequency error, most possibly because the confidence of an association rule is the ratio between the frequencies of two itemsets that appear in the same transactions and their sample frequencies will therefore have similar errors that cancel out when the ratio is computed. Similar conclusions can be made for the absolute  $\varepsilon$ -close approximation case (not reported due to space constraints).

The major motivating intuition for the use of sampling in market basket analysis tasks is that mining a sample of the dataset is faster than mining the entire dataset. Nevertheless, the mining time does not only depend on the number of transactions, but also on the number of frequent itemsets. Given that our methods suggest to mine the sample at a lowered minimum frequency threshold, this may cause an increase in running time that would make our method not useful in practice, because there may be many more frequent itemsets than at the original frequency threshold. We performed a number of experiments to evaluate whether this was the case and present the results in Fig. 3. We mined the artificial dataset introduced before for different values of  $\theta$ , and created samples of size sufficient to obtain a relative  $\varepsilon$ -close approximation to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$ , for  $\varepsilon = 0.05$  and  $\delta = 0.1$ . Figure 3 shows the time needed to mine the large dataset and the time needed to create and mine the samples. It is possible to appreciate that, even considering the sampling time, the speed up achieved by our method is relevant, proving the usefulness of sampling.

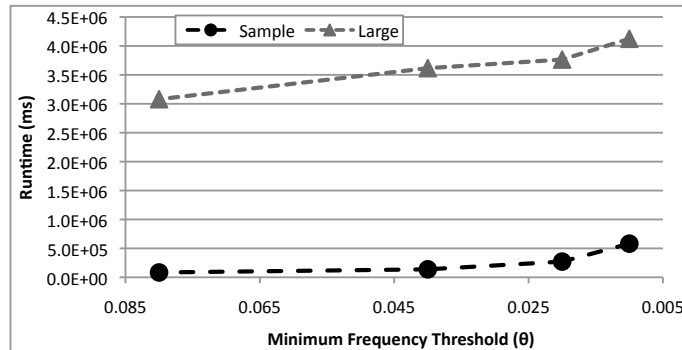


Figure 3: Runtime Comparison. The sample line includes the sampling time.

Comparing our results to previous work we note that the bounds generated by our technique are always linear in the VC-dimension  $d$  associated with the dataset. As reported in Table 1, the best previous work [6] presented bounds that are linear in the maximum transaction size  $\Delta$  for two of the six problems studied here. Figures 4a and 4b shows a comparison of the actual sample sizes for relative  $\varepsilon$ -close approximations to  $\text{FI}(\mathcal{D}, \mathcal{I}, \theta)$  for as function of  $\theta$  and  $\varepsilon$ . To compute the points for these figures, we set  $\Delta = d = 50$ ,

corresponding to the worst possible case for our method, i.e., when the VC-dimension of the range space associated to the dataset is exactly equal to the maximum transaction length. We also fixed  $\delta = 0.05$  (the two methods behave equally as  $\delta$  changes). For Fig. 4a, we fixed  $\varepsilon = 0.05$ , while for Fig. 4b we fixed  $\theta = 0.05$ . From the Figures we can appreciate that both bounds have similar, but not equal, dependencies on  $\theta$  and  $\varepsilon$ . More precisely the bound presented in this work is less dependent on  $\varepsilon$  and only slightly more dependent on  $\theta$ . It also evident that the sample sizes suggested by the bound presented in this work are always much smaller than those presented in [6] (the vertical axis has logarithmic scale). In this comparison we used  $\Delta = d$ , but almost all real datasets we encountered have  $d \ll \Delta$  as shown in Table 2 which would result in a larger gap between the sample sizes provided by the two methods.

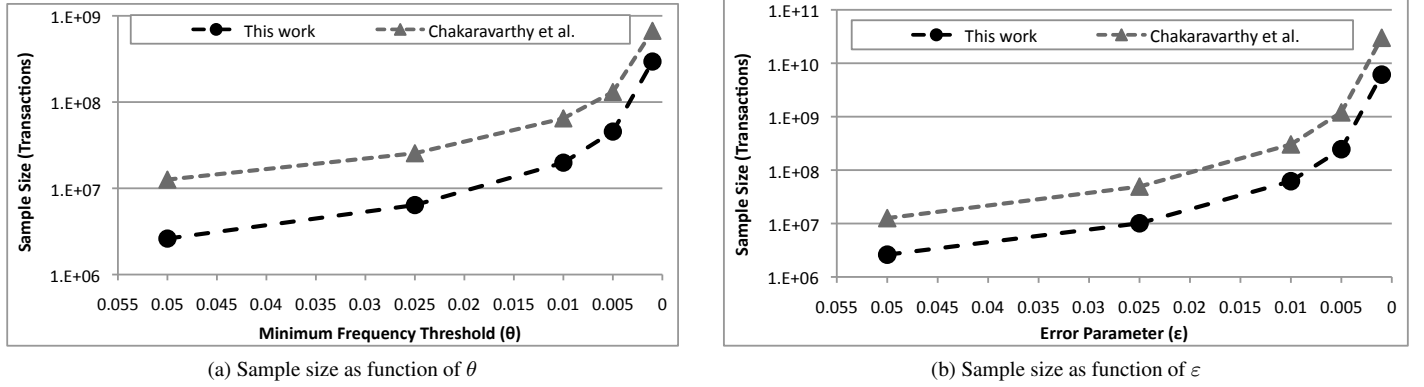


Figure 4: Sample sizes for relative  $\varepsilon$ -close approximations to  $FI(\mathcal{D}, \mathcal{I}, \theta)$ .

	accidents	BMS-POS	BMS-Webview-1	kosarak	mushroom	pumsb*	retail	webdocs
$\Delta$	51	164	267	2497	23	63	76	71472
$d$	46	81	57	443	23	59	58	2452

Table 2: Values for maximum transaction length  $\Delta$  and d-index  $d$  for real datasets

## 6 Conclusions

In this paper we presented a novel technique to derive random sample sizes sufficient to easily extract high-quality approximations of the (top- $K$ ) frequent itemsets and of the collection of association rules. The sample size are linearly dependent on the VC-Dimension of the range space associated to the dataset, which is upper bounded by the maximum integer  $d$  such that there at least  $d$  transactions of length at least  $d$  in the dataset. This bound is tight for a family of datasets. We conducted an extensive experimental evaluation which shows the practical usefulness of our method, confirming our theoretical analysis. In the future we would like to explore possible ways of giving a stricter upper bound to the VC-dimension for a given dataset, or whether other measures of sample complexity like the triangular rank [29] can suggest smaller samples sizes.

## References

- [1] AGRAWAL, R., IMIELIŃSKI, T., AND SWAMI, A. 1993. Mining association rules between sets of items in large databases. *SIGMOD Rec.* 22, 207–216.
- [2] AGRAWAL, R. AND SRIKANT, R. 1994. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.
- [3] ALON, N. AND SPENCER, J. H. 2008. *The Probabilistic Method* third Ed. Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Hoboken, NJ, USA.

- [4] BRÖNNIMANN, H., CHEN, B., DASH, M., HAAS, P., AND SCHEUERMANN, P. 2003. Efficient data reduction with ease. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '03. ACM, New York, NY, USA, 59–68.
- [5] CEGLAR, A. AND RODDICK, J. F. 2006. Association mining. *ACM Comput. Surv.* 38, 5.
- [6] CHAKARAVARTHY, V. T., PANDIT, V., AND SABHARWAL, Y. 2009. Analysis of sampling techniques for association rule mining. In *Proceedings of the 12th International Conference on Database Theory*. ICDT '09. ACM, New York, NY, USA, 276–283.
- [7] CHANDRA, B. AND BHASKAR, S. 2011. A new approach for generating efficient sample from market basket data. *Expert Systems with Applications* 38, 3, 1321 – 1325.
- [8] CHAZELLE, B. 2000. *The discrepancy method: randomness and complexity*. Cambridge University Press, New York, NY, USA.
- [9] CHEN, B., HAAS, P., AND SCHEUERMANN, P. 2002. A new two-phase sampling based algorithm for discovering association rules. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '02. ACM, New York, NY, USA, 462–468.
- [10] CHEN, C., HORNG, S.-J., AND HUANG, C.-P. 2011. Locality sensitive hashing for sampling-based algorithms in association rule mining. *Expert Systems with Applications* 38, 10, 12388 – 12397.
- [11] CHEUNG, Y.-L. AND FU, A. W.-C. 2004. Mining frequent itemsets without support threshold: With and without item constraints. *IEEE Trans. on Knowl. and Data Eng.* 16, 1052–1069.
- [12] CHUANG, K.-T., CHEN, M.-S., AND YANG, W.-C. 2005. Progressive sampling for association rules based on sampling error estimation. In *Advances in Knowledge Discovery and Data Mining*, T. Ho, D. Cheung, and H. Liu, Eds. Lecture Notes in Computer Science Series, vol. 3518. Springer, Berlin / Heidelberg, 37–44.
- [13] CHUANG, K.-T., HUANG, J.-L., AND CHEN, M.-S. 2008. Power-law relationship and self-similarity in the itemset support distribution: analysis and applications. *The VLDB Journal* 17, 5, 1121–1141.
- [14] FU, A. W.-C., KWONG, R. W.-W., AND TANG, J. 2000. Mining n-most interesting itemsets. In *Proceedings of the 12th International Symposium on Foundations of Intelligent Systems*. ISMIS '00. Springer, Berlin / Heidelberg, 59–67.
- [15] HAN, J., CHENG, H., XIN, D., AND YAN, X. 2007. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery* 15, 55–86.
- [16] HAR-PELED, S. AND SHARIR, M. 2011. Relative  $(p, \varepsilon)$ -approximations in geometry. *Discrete & Computational Geometry* 45, 3, 462–496.
- [17] HU, X. AND YU, H. 2006. The research of sampling for mining frequent itemsets. In *Rough Sets and Knowledge Technology*, G.-Y. Wang, J. Peters, A. Skowron, and Y. Yao, Eds. Lecture Notes in Computer Science Series, vol. 4062. Springer, Berlin / Heidelberg, 496–501.
- [18] HWANG, W. AND KIM, D. 2006. Improved association rule mining by modified trimming. In *Proceedings of the 6th IEEE International Conference on Computer and Information Technology*. CIT '06. IEEE Computer Society, 24.
- [19] JIA, C. AND LU, R. 2005. Sampling ensembles for frequent patterns. In *Fuzzy Systems and Knowledge Discovery*, L. Wang and Y. Jin, Eds. Lecture Notes in Computer Science Series, vol. 3613. Springer, Berlin / Heidelberg, 478–478.
- [20] JIA, C.-Y. AND GAO, X.-P. 2005. Multi-scaling sampling: An adaptive sampling method for discovering approximate association rules. *Journal of Computer Science and Technology* 20, 309–318.
- [21] JOHN, G. H. AND LANGLEY, P. 1996. Static versus dynamic sampling for data mining. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD '96. The AAAI Press, Menlo Park, CA, USA, 367–370.
- [22] LI, Y. AND GOPALAN, R. 2005. Effective sampling for mining association rules. In *AI 2004: Advances in Artificial Intelligence*, G. Webb and X. Yu, Eds. Lecture Notes in Computer Science Series, vol. 3339. Springer, Berlin / Heidelberg, 73–75.
- [23] LI, Y., LONG, P. M., AND SRINIVASAN, A. 2001. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences* 62, 3, 516–527.

- [24] LINIAL, N., MANSOUR, Y., AND RIVEST, R. L. 1991. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation* 90, 1, 33–49.
- [25] LÖFFLER, M. AND PHILLIPS, J. M. 2009. Shape fitting on point sets with probability distributions. In *Algorithms - ESA 2009*, A. Fiat and P. Sanders, Eds. Lecture Notes in Computer Science Series, vol. 5757. Springer, Berlin / Heidelberg, 313–324.
- [26] MAHAFAZAH, B. A., AL-BADARNEH, A. F., AND ZAKARIA, M. Z. 2009. A new sampling technique for association rule mining. *Journal of Information Science* 35, 3, 358–376.
- [27] MAMPAEY, M., TATTI, N., AND VREEKEN, J. 2011. Tell me what i need to know: succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11. ACM, New York, NY, USA, 573–581.
- [28] MANNILA, H., TOIVONEN, H., AND VERKAMO, I. 1994. Efficient algorithms for discovering association rules. In *KDD Workshop*. The AAAI Press, Menlo Park, CA, USA, 181–192.
- [29] NEWMAN, I. AND RABINOVICH, Y. 2012. On multiplicative  $\lambda$ -approximations and some geometric applications-approximations and some geometric applications. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '12. SIAM, 51–67.
- [30] PARTHASARATHY, S. 2002. Efficient progressive sampling for association rules. In *Proceedings of the 2002 IEEE International Conference on Data Mining*. ICDM '02. IEEE Computer Society, 354–361.
- [31] PIETRACAPRINA, A., RIONDATO, M., UPFAL, E., AND VANDIN, F. 2010. Mining top-K frequent itemsets through progressive sampling. *Data Mining and Knowledge Discovery* 21, 310–326.
- [32] PIETRACAPRINA, A. AND VANDIN, F. 2007. Efficient incremental mining of top-K frequent closed itemsets. In *Discovery Science*, V. Corruble, M. Takeda, and E. Suzuki, Eds. Lecture Notes in Computer Science Series, vol. 4755. Springer, Berlin / Heidelberg, 275–280.
- [33] SCHEFFER, T. AND WROBEL, S. 2002. Finding the most interesting patterns in a database quickly by using sequential sampling. *J. Mach. Learn. Res.* 3, 833–862.
- [34] TOIVONEN, H. 1996. Sampling large databases for association rules. In *Proceedings of the 22th International Conference on Very Large Data Bases*. VLDB '96. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 134–145.
- [35] VAPNIK, V. N. 1999. *The Nature of Statistical Learning Theory*. Statistics for engineering and information science. Springer-Verlag, New York, NY, USA.
- [36] VAPNIK, V. N. AND CHERVONENKIS, A. J. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 16, 2, 264–280.
- [37] VASUDEVAN, D. AND VOJONOVIĆ, M. 2009. Ranking through random sampling. MSR-TR-2009-8 8, Microsoft Research.
- [38] WANG, J., HAN, J., LU, Y., AND TZVETKOV, P. 2005. TFP: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans. on Knowl. and Data Eng.* 17, 652–664.
- [39] WANG, S., DASH, M., AND CHIA, L.-T. 2005. Efficient sampling: Application to image data. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD 2005*, T. B. Ho, D. W.-L. Cheung, and H. Liu, Eds. Lecture Notes in Computer Science Series, vol. 3518. Springer, Berlin/ Heidelberg, 452–463.
- [40] ZAKI, M., PARTHASARATHY, S., LI, W., AND OGIHARA, M. 1997. Evaluation of sampling for data mining of association rules. In *Proceedings of the Seventh International Workshop on Research Issues in Data Engineering*. RIDE '97. IEEE Computer Society, 42–50.
- [41] ZHANG, C., ZHANG, S., AND WEBB, G. I. 2003. Identifying approximate itemsets of interest in large databases. *Applied Intelligence* 18, 91–104.
- [42] ZHAO, Y., ZHANG, C., AND ZHANG, S. 2006. Efficient frequent itemsets mining by sampling. In *Proceeding of the 2006 conference on Advances in Intelligent IT: Active Media Technology 2006*. IOS Press, Amsterdam, The Netherlands, 112–117.